SDGAN: Disentangling Semantic Manipulation for Facial Attribute Editing

Wenmin Huang[1], **Weiqi Luo**[1*], **Jiwu Huang**[2], **Xiaochun Cao**[3]

[1] School of Computer Science and Engineering, Sun Yat-sen University, China
[2] Shenzhen Key Laboratory of Media Security, Shenzhen University, China
[3] School of Cyber Science and Technology, Sun Yat-sen University, China
huangwm36@mail2.sysu.edu.cn, {luoweiqi, caoxiaochun}@mail.sysu.edu.cn, jwhuang@szu.edu.cn

Abstract

Facial attribute editing has garnered significant attention, yet prevailing methods struggle with achieving precise attribute manipulation while preserving irrelevant details and controlling attribute styles. This challenge primarily arises from the strong correlations between different attributes and the interplay between attributes and identity. In this paper, we propose Semantic Disentangled GAN (SDGAN), a novel method addressing this challenge. SDGAN introduces two key concepts: a semantic disentanglement generator that assigns facial representations to distinct attribute-specific editing modules, enabling the decoupling of the facial attribute editing process, and a semantic mask alignment strategy that confines attribute editing to appropriate regions, thereby avoiding undesired modifications. Leveraging these concepts, SDGAN demonstrates accurate attribute editing and achieves high-quality attribute style manipulation through both latent-guided and reference-guided manners. We extensively evaluate our method on the CelebA-HQ database, providing both qualitative and quantitative analyses. Our results establish that SDGAN significantly outperforms state-of-the-art techniques, showcasing the effectiveness of our approach. The code implementing our model is available at https://github.com/sysuhuangwenmin/SDGAN.

Introduction

Facial attribute editing aims to modify facial images by altering specific attributes, finding applications in diverse real-world domains, including entertainment, visual effects, and e-commerce. With the rapid development of deep generative models, facial attribute editing powered with generative adversarial networks (GANs) (Goodfellow et al. 2014) has achieved impressive progress (He et al. 2019; Lee et al. 2020; Dalva, Altındiş, and Dundar 2022; Shi et al. 2022).

The challenges in facial attribute editing primarily arise from two main aspects: 1) Correct Modification and Unrelated Preservation: An effective model should accurately manipulate the target attribute while preserving irrelevant details such as non-target attributes, identity information, and illumination, as illustrated in the 1st row in Fig. 1. Achieving this ideal facial attribute editing is difficult due to the strong correlation between different attributes (e.g., eyeglasses and age) and between attributes and identity; 2) Style Manipulation: Adding or removing attributes have limited real-world applicability. In many cases, there is a need to manipulate the style of the target attribute, such as different styles of eyeglasses: sunglasses, myopic glasses, and reading glasses, as shown in the 2nd row in Fig. 1. However, this task poses particular challenges as labeled databases with diverse attribute styles are often unavailable.

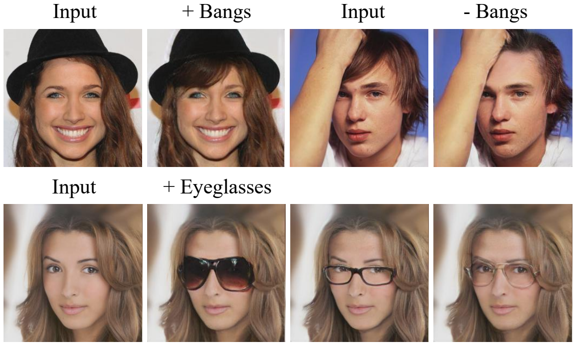


Figure 1: Illustration of correct modification and unrelated preservation (the 1st row) & style manipulation (the 2nd row) with the proposed SDGAN.

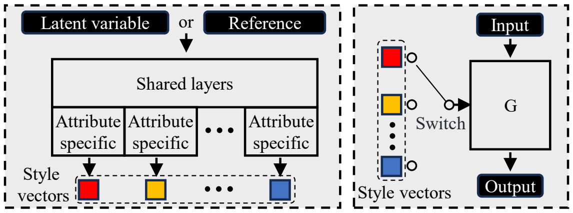
Numerous related approaches have been proposed to achieve accurate modifications while preserving unrelated attributes. These methods can be broadly categorized into two groups: image-to-image translation and latent space manipulation, based on differences in their underlying principles. Image-to-image translation methods typically employ an encoder-decoder architecture to directly learn the translation between the original and edited images. Early works (Shen et al., 2017; Zhang et al., 2018) use cycle-consistent and adversarial losses to train different models for various attributes. More recent studies (Liu et al., 2019; Chen et al., 2020) have adopted conditional Generative Adversarial Networks (cGANs) and used attribute labels as generative conditions, enabling a single model to handle multiple attribute editing tasks; On the other hand, those methods based on latent space manipulation (Wang et al., 2022; Pehlivan et al., 2023) utilize GAN inversion to obtain the latent representation of a given face. They then manipulate this latent representation along semantic directions and feed it into a

[*] Corresponding Author.

pre-trained GAN, such as StyleGAN (Karras et al. 2020), to generate the edited results. While using pre-trained GANs on large databases, such as FFHQ (Karras, Laine, and Aila 2019), can yield realistic editing results, it often sacrifices fine details in the original image due to the challenges of jointly training GAN inversion and pre-trained GANs. Furthermore, both these approaches are unable to achieve attribute style manipulation, which restricts their applicability in real-world scenarios. Future research may focus on addressing this limitation and exploring methods that can perform attribute style manipulation while preserving fine details in the edited images.

Several studies have independently explored two approaches for style manipulation. The first approach, known as the latent-guided manner, involves embedding latent variables sampled from Gaussian noise into the generator (Wang et al. 2019). The second approach, referred to as the reference-guided manner, utilizes style vectors extracted from reference images (Xiao, Hong, and Ma 2018). Recent studies (Choi et al. 2020; Li et al. 2021b; Dalva, Altındiş, and Dundar 2022) have attempted to combine both types of style manipulation into a unified framework. In this unified approach, researchers utilize distinct style extraction modules for randomly sampled Gaussian noise and reference images, and style vectors for each attribute are independently generated through the different output branches of these modules. This facial attribute editing paradigm has achieved impressive results for style manipulation. However, it is worth noting that the attribute editing process in the generator is not decomposed, so there is a possibility that the approach could fail to accurately manipulate the target attribute and may inadvertently preserve irrelevant details, as observed in (Choi et al. 2020).

In this paper, we present an innovative framework called Semantic Disentangled GAN (SDGAN) to tackle the challenges in facial attribute editing. Building upon prior research (Choi et al. 2020; Li et al. 2021b), SDGAN initially utilizes similar modules to create style vectors for all attributes based on latent variables and reference images. Subsequently, SDGAN introduces two innovative concepts for face attribute editing: a semantic disentanglement generator and a semantic mask alignment strategy. Semantic disentanglement generator factors the latent representation of SDGAN by the attribute-specific editing modules and semantic masks. Each attribute-specific editing module is individually modulated with its corresponding attribute style vector, and an image is synthesized by composing local feature maps with semantic masks. Unlike generators using a shared editing module (Choi et al. 2020) or modular editing (Zhao et al. 2018; Li et al. 2021b), our method utilizes semantic masks to explicitly divide the manipulation regions of different attribute-specific editing modules, thus effectively decoupling the editing process of face attributes. Furthermore, to achieve correct modification and unrelated preservation, semantic masks should focus on the semantic regions corresponding to the target attribute. To this end, we design a simple yet effective semantic mask alignment strategy. This strategy only requires attribute labels for training, without the need for cumbersome semantic region annotation for target attribute. Its goal is to force that the focus region for the semantic masks supports the attribute classifier in making correct predictions regarding the target attribute.



(a) Style Disentanglement Extractor (b) Generator

Figure 2: Illustration of (a) style disentanglement extractor and (b) generator for facial attribute editing.

In summary, our contributions are as follows:

- We introduce a semantic disentanglement generator that effectively factors the facial attribute editing process by attribute-specific editing modules and semantic masks.
- We propose a simple yet effective semantic mask alignment strategy to confine attribute editing within the appropriate regions.
- Extensive comparative experiments show that our method achieves state-of-the-art results across various metrics, both qualitatively and quantitatively.

Related Work

Generative Adversarial Networks (GANs). GANs (Goodfellow et al. 2014) have dominated the facial attribute editing due to their remarkable image synthesis capabilities. Consisting of a generator and a discriminator, GANs are trained in an adversarial manner to learn the distribution of real images. Recently, substantial efforts have been devoted to enhance the training stability (Petzka, Fischer, and Lukovnikov 2018; Yazici et al. 2019) and image synthesis quality (Brock, Donahue, and Simonyan 2019; Karras, Laine, and Aila 2019) of GANs, resulting in their widespread use as the dominant model for various image-to-image translation tasks. Such tasks include image inpainting (Li et al. 2021a), object insertion (Gafni and Wolf 2020), image super-resolution (Zhang et al. 2019; Xin et al. 2020) colorization (Isola et al. 2017), image editing (Xu et al. 2021; Xia et al. 2023), and more.

Facial attribute editing. Up to now, many facial attribute editing methods based on image-to-image translation (Shen and Liu 2017; Yin, Liu, and Loy 2019; Gao et al. 2021) and latent space manipulation (Wang et al. 2022; Alaluf et al. 2022; Pehlivan, Dalva, and Dundar 2023) have demonstrated impressive results in accurate modification and unrelated preservation. However, these approaches overlook the importance of attribute style manipulation in real-world scenarios. Recent studies have attempted to incorporate attribute style manipulation into facial attribute editing. For example, ELEGANT (Xiao, Hong, and Ma 2018) introduces a reference-guided style manipulation approach. It first employs an encoder to obtain latent encodings for both the original and reference images. Then, by exchanging the target

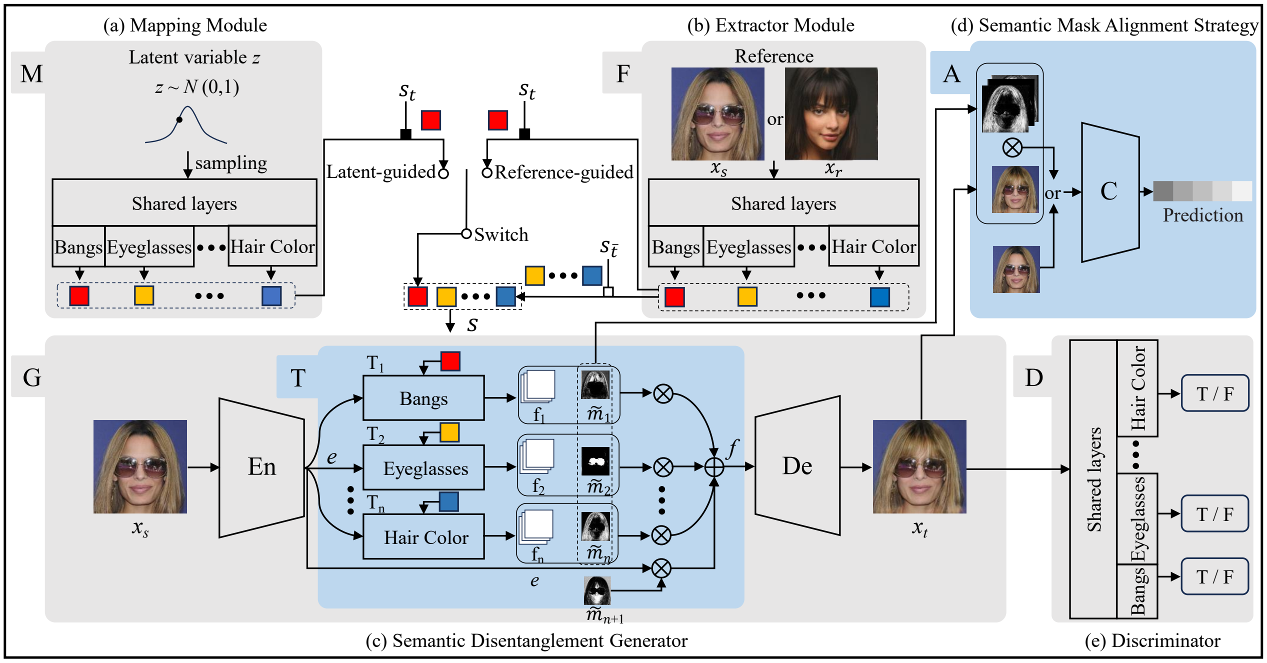


Figure 3: Overview of the proposed SDGAN, comprising the following components: (a) mapping model M, (b) extractor module F, (c) semantic disentanglement generator G, (d) semantic mask alignment strategy A, and (e) discriminator D. $\otimes$ and $\oplus$ denote element-wise multiplication and addition operations, respectively.

attribute part of their latent encodings, ELEGANT transfers the target attribute style from the reference image to the input image. SDIT (Wang et al. 2019) explores a latent-guided style manipulation approach. It is built upon a conditional encoder-generator framework and achieves both correctness of attribute editing and diversity in the output style by simultaneously using attribute labels and latent variables sampled from Gaussian noise as generation conditions. Subsequently, StarGANv2 (Choi et al. 2020), HiSD (Li et al. 2021b) and VecGAN (Dalva, Altındiş, and Dundar 2022) combine the above two types of style manipulation approaches into a unified framework. These methods follow a similar editing pipeline, as shown in Fig. 2. First, the style extractor with multiple output branches generates style vectors from latent variable or reference image. Then, the generator utilizes target labels to retrieve corresponding style vector for both target attribute editing and style manipulation. In this paper, we employ a style extractor and mapping network similar to HiSD to leverage its excellent style manipulation capabilities. However, we introduce two novel components: a semantic disentanglement generator (Fig. 3 (c)) and a semantic mask alignment strategy (Fig. 3 (d)). These additions are aimed at improving the editing quality and accuracy, and we will elaborate on these new issues in the following section.

Approach

The primary idea behind our SDGAN is to achieve precise attribute editing and style manipulation by decoupling the facial attribute editing process and constraining attribute editing region. As shown in Fig. 3, the SDGAN framework consists of five components: the mapping module M, the extractor module F, the semantic disentanglement generator G, the semantic mask alignment strategy A, and the discriminator D. These components work in unison to generate, extract, and manipulate the style of the target attributes in facial images. In the subsequent sections, we provide a comprehensive description of key components in SDGAN, including modules M, F, as well as the novel components G and A. Subsequently, we delve into the discussion of our objective functions.

Mapping and Extractor Modules

The modules M and F are used to generate and extract initial style vectors from Gaussian noise and reference images, respectively. Both M and F have separate output branches for each attribute. Depending on whether the target attribute style is provided by M or F, there are two ways to get the final style vectors for manipulation: latent-guided manner and reference-guided manner.

1) In the latent-guided manner, given an facial image x_s to be edited and a target attribute index t (e.g., bangs or eyeglasses), M generates the target attribute style vector s_t from randomly sampled Gaussian noise z. To preserve irrelevant attributes unchanged, F extracts the target attribute-irrelevant style vectors $s_{\bar{t}} = \{s_i | \forall i \neq t\}_{i=1...n}$ from the reference image x_s, where s_i indicates the style vector of i-th attribute, and n is the number of attributes to be considered. Finally, we combine s_t and $s_{\bar{t}}$ to form the final style vectors $s = \{s_1, s_2, ..., s_t, ..., s_n\}$.

2) In the reference-guided manner, the extractor module F is responsible for extracting the target attribute style vector s_t and the target attribute-irrelevant vectors $s_{\bar{t}}$ from a given reference image x_r and the input image x_s, respectively. By combining s_t with $s_{\bar{t}}$, we create the final style vectors s. It's important to note that in this process, we do not utilize module M at all. The primary objective here is to transfer the target attribute style from x_r to x_s.

Semantic Disentanglement Generator

To effectively manipulate the target attribute while preserving other irrelevant details, achieving a strong disentanglement of face representations is crucial. Previous methods (Choi et al. 2020; Li et al. 2021b) have achieved impressive results by decoupling the generation and extraction processes of style vectors. However, their generators mainly focus on obtaining manipulation results that reflect the target attribute style vector, ignoring other unrelated attributes and content, as illustrated in Fig. 2 (b). In contrast, our approach takes all attribute style vectors into account as generation conditions, enabling us to manipulate the target attribute style while preserving other unrelated attributes and content. Moreover, our generator is designed to learn distinct attribute-specific editing modules and editing regions for different attributes, effectively decoupling the editing process of facial attributes.

The generator G aims to transform the input x_s into the edited image x_t based on the resulting style vectors s derived from the mapping and extractor modules, denoted as $x_t = G(x_s, s)$. As shown in Fig. 3 (c), G is composed of three parts, i.e., the encoder En, the transformer T, and the decoder De. En first converts x_s into its immediate feature e. T then assigns e to different attribute-specific editing modules T_i ($i = 1, 2, ..., n$). Each T_i utilizes adaptive instance normalization (AdaIN) (Huang and Belongie 2017) to inject the style vector s_i into e, enabling it to learn the editing for the i-th attribute. This process generates an edited feature f_i and a semantic mask m_i that is used to restrict the manipulation region of T_i. To encourage that each T_i focuses solely on its corresponding attribute and minimizes undesired changes, we employ Softmax to divide the manipulation region of each T_i:

$$\tilde{m}_i = \frac{\exp(m_i)}{\sum_{i'=1}^{n+1}\exp(m_{i'})}, i = 1, 2, \dots n+1, \tag{1}$$

where m_{n+1} is simply set to as a zero matrix, capturing the region not associated with any specific attribute. Moreover, the feature maps from different T_i are aggregated as follows:

$$f = \sum_{i=1}^{n} f_i \otimes \tilde{m}_i + e \otimes \tilde{m}_{n+1}, \tag{2}$$

where $\otimes$ indicates element-wise multiplication operation. Finally, f is fed into De to generate the edited image x_t.

Semantic Mask Alignment Strategy

In the previous section, we introduce the semantic mask $\tilde{m}_i$, which aims to facilitate correct attribute modification while suppressing undesired changes. However, accurately identifying the relevant semantic regions associated with the i-th attribute is a challenging task, especially without using additional supervised signals. To address this challenge, we first propose an attribute classifier C to predict the attributes in the image x_s, and we define the corresponding loss function of C as follows:

$$\mathcal{L}_c = \sum_{i=1}^{n} -a_i \log C_i(x_s) - (1 - a_i)\log(1 - C_i(x_s)), \tag{3}$$

where $a_i \in \{0, 1\}$ denotes the true label of the i-th attribute of x_s, while $C_i(x_s) \in [0, 1]$ denotes the corresponding prediction with the classifier C. As for $\tilde{m}_i$, we force it to support the C to accurately predict the label $b_i \in \{0, 1\}$ of i-th attribute of x_t by using dynamic weighted cross-entropy:

$$\begin{aligned}\mathcal{L}_m = \sum_{i=1}^{n} -\frac{1}{\overline{m}_i + \epsilon}[b_i C_i(x_t \otimes \tilde{m}_i)+ \\ (1 - b_i)(1 - C_i(x_t \otimes \tilde{m}_i))],\end{aligned} \tag{4}$$

where $\overline{m}_i$ is the mean value of the mask $\tilde{m}_i$, which is used to enhance training stability, and ϵ is a hyperparameter. $\mathcal{L}_c$ and $\mathcal{L}_m$ respectively optimize C and G by iteration way.

Training Objectives

Reconstruction objective. To encourage the editing results to preserve as many details as possible from the input image, we consider three reconstruction versions:

1) $x'_s = De(En(x_s))$: When no attribute editing operation is involved, the output should faithfully reconstruct the input image.
2) $x''_s = G(x_s, F(x_s))$: Under the guidance of the style vectors of the input image, the output of G should reconstruct the input image.
3) $x'''_s = G(G(x_s, s), F(x_s))$: We first edit x_s into $x_t = G(x_s, s)$ using latent-guided manner. Then, taking x_s as the reference image, we reverse x_t back to $x'''_s = G(x_t, F(x_s))$ using reference-guided manner.

We define the reconstruction objective as follows:

$$\mathcal{L}_{rec} = ||x'_s - x_s||_1 + ||x''_s - x_s||_1 + ||x'''_s - x_s||_1. \tag{5}$$

Adversarial objective. To promote realistic editing, D is designed with multiple output branches, and each branch learns to distinguish whether an image is a real image of the corresponding attribute or a image generated by G:

$$\mathcal{L}_{adv} = \log D_t(x_s) + \log(1 - D_t(x_t)), \tag{6}$$

where D_t indicates the output branch corresponding to the target attribute t. This objective encourages M to generate the correct target attribute style and compels F to accurately extract the target attribute style from the reference image.

Style objective. The style vectors extracted from the edited image x_t is supposed to be equal to the style vectors s injected into G, which is defined as:

$$\mathcal{L}_{sty} = ||F(x_t) - s||_1. \tag{7}$$

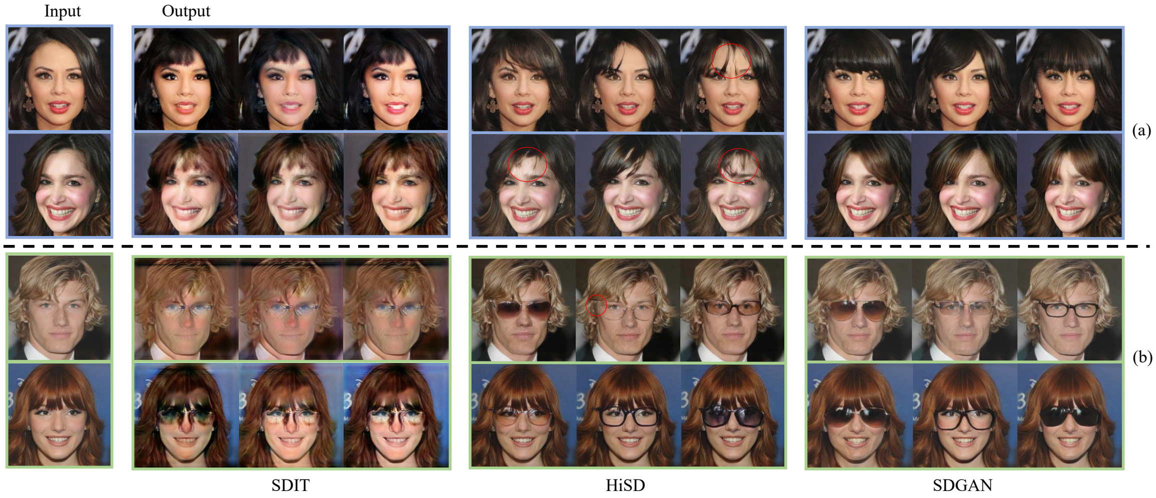


Figure 4: Qualitative results of the latent-guided manner, including (a) + Bangs and (b) + Eyeglasses.

Method	+ Bangs		- Bangs		+ Eleglasses		- Eleglasses		Average	
	↓ FID	↑ Acc	↓ FID	↑ Acc	↓ FID	↑ Acc	↓ FID	↑ Acc	↓ FID	↑ Acc
SDIT	23.68	**92.44**	41.33	96.93	77.65	90.28	107.27	98.55	62.48	94.55
VecGAN	20.17	-	-	-	-	-	-	-	-	-
HiSD	17.18	85.99	39.71	92.25	62.17	80.39	95.99	**99.33**	53.76	89.49
Ours	**15.69**	90.67	**38.69**	**97.67**	**49.39**	**98.52**	**78.99**	95.33	**45.69**	**95.54**

Table 1: Quantitative results of the latent-guided manner.

This objective encourages G to utilize s during the generation of x_t, enabling F to accurately extract s from x_t.

Full objective. Finally, our full objective function is written as follows:

$$\min_{M,F,G,C}\max_{D}(\mathcal{L}_c + \lambda_m\mathcal{L}_m + \lambda_{rec}\mathcal{L}_{rec} + \mathcal{L}_{adv} + \lambda_{sty}\mathcal{L}_{sty}), \tag{8}$$

where λ_m, λ_{rec}, and λ_{sty} are hyperparameters for each term.

Experiments

Database. Like previous methods (Li et al. 2021b; Pehlivan, Dalva, and Dundar 2023), we evaluate our method on CelebA-HQ (Karras et al. 2018), which comprises 30,000 facial images with attribute annotations. Following (Li et al. 2021b), we split CelebA-HQ into a test set of 3,000 images and a training set of 27,000 images. For the editing task, we focus on three typical attributes: bangs, eyeglasses, and hair color, which are commonly considered in existing methods.

Baselines. Since the proposed method can control attribute styles, four related methods are included for comparative study, including SDIT (Wang et al. 2019), which supports only the latent-guided manner, ELEGANT (Xiao, Hong, and Ma 2018), which supports only the reference-guided manner, HiSD (Li et al. 2021b), and VecGAN (Dalva, Altındiş, and Dundar 2022). Note that for VecGAN, we provide only quantitative results as its code is not available. In addition, we also compare our method with two modern methods (i.e., HFGI (Wang et al. 2022) and StyleRes (Pehlivan, Dalva, and Dundar 2023)) based on latent space manipulation, since these methods usually achieve high-fidelity editing results.

Evaluation metrics. We evaluate both the visual quality of generated images and the attribute editing ability using Frechet inception distance (FID (Heusel et al. 2017)) and attribute editing accuracy (Acc) for our method and baselines. Following the evaluation protocol (Li et al. 2021b), we report FID for the attribute of bangs, and consider challenging eyeglasses as an additional reference indicator. For Acc, we use an attribute classifier trained on the training set of CelebA-HQ, achieving an accuracy of 95.0% on the test set.

Comparison with Style Manipulation Methods

In this section, we evaluate the style manipulation performance of our method from two perspectives: latent-guided manner and reference-guided manner.

Latent-guided manner. Fig. 4 shows qualitative results of the competing methods. SDIT exhibits limited performance in style manipulation and image quality. HiSD produces produces unrealistic bangs (Fig. 4 (a)) and eyeglasses (Fig. 4 (b)). In contrast, our method effectively generates high-quality and realistic editing results. Quantitative results are listed in Table 1. Our approach achieves the best FID and average Acc. Most significantly, our method demonstrates significant improvements in average FID compared to HiSD, with a decrease of 8.07. For hair color manipulation and additional results, please refer to our supplemental material.

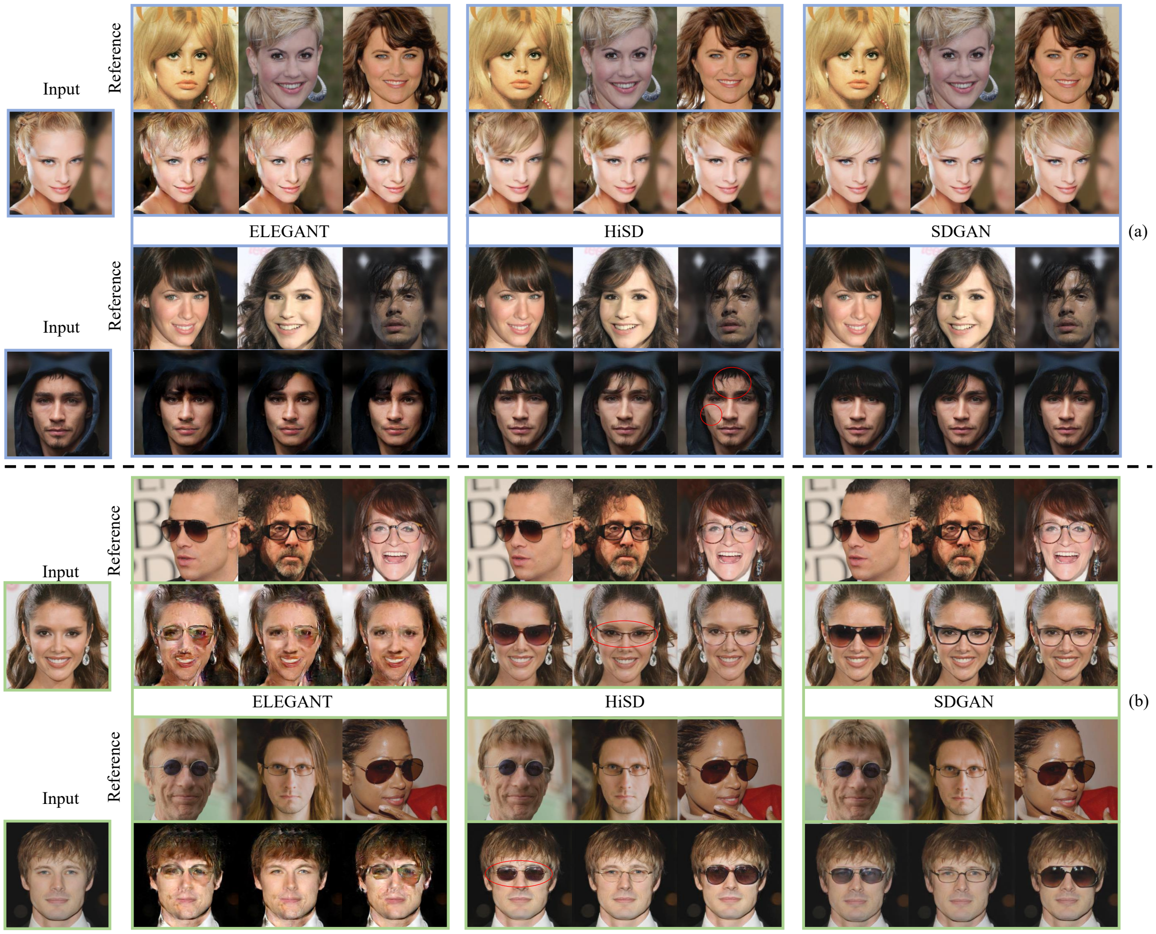


Figure 5: Qualitative results of the reference-guided manner, including (a) + Bangs and (b) + Eyeglasses.

Method	+ Bangs		- Bangs		+ Eleglasses		- Eleglasses		Average	
	↓ FID	↑ Acc	↓ FID	↑ Acc	↓ FID	↑ Acc	↓ FID	↑ Acc	↓ FID	↑ Acc
ELEGANT	28.16	72.88	52.11	85.28	91.10	68.27	107.08	**98.66**	69.61	81.27
VecGAN	20.72	-	-	-	-	-	-	-	-	-
HiSD	17.53	76.32	38.73	89.33	56.64	83.82	89.97	95.11	50.71	86.14
Ours	**16.27**	**86.27**	**38.62**	**90.52**	**51.45**	**87.09**	**75.17**	95.02	**45.37**	**89.70**

Table 2: Quantitative results of the reference-guided manner.

Reference-guided manner. Fig. 5 presents qualitative results of the competing methods. ELEGANT fails to preserve irrelevant content when manipulating eyeglasses. In certain cases, HiSD cannot accurately transfer the shape and color of the eyeglasses, as evident in the middle column of the second row, and the left column of the last row in Fig. 5 (b). In contrast, our method accurately transfers the attribute style of reference images and obtain high-quality editing results. Quantitative results are listed in Table 2. Our approach consistently achieves the best FID and average Acc, and demonstrates significant improvements in average FID compared to HiSD, with a decrease of 5.34.

Comparison with Methods Based on Latent Space Manipulation

In this section, our main focus is to compare the challenging attribute of eyeglasses. Note that both HFGI and StyleRes utilize StyleGAN as the generator and have been pre-trained on FFHQ. However, they tend to suffer from the loss of fine details in the original images due to the absence of end-to-end training, as evident in Fig. 6. For example, they encounter difficulties in preserving the details of the collar (1st row) and introduce distortions in the earrings (3th row). Additionally, they struggle to handle occlusions effectively (2nd row). In contrast, our method effectively preserves image details while achieving accurate attribute editing.

Method	+ Bangs		- Bangs		+ Eleglasses		- Eleglasses		Average	
	L	R	L	R	L	R	L	R	L	R
w/o ASEM	18.41	18.38	38.75	42.97	50.45	52.91	80.08	76.89	46.92	47.78
w/o Mask	18.43	18.32	41.20	41.27	50.93	52.85	83.03	79.50	48.39	47.98
w/o SMAS	17.26	17.93	41.31	40.90	54.72	61.54	87.88	86.56	50.29	51.73
SDGAN	**15.69**	**16.27**	**38.69**	**38.62**	**49.39**	**51.45**	**78.99**	**75.17**	**45.69**	**45.37**

Table 3: Quantitative results of the ablation study on FID. L: latent-guided; R: reference-guided.

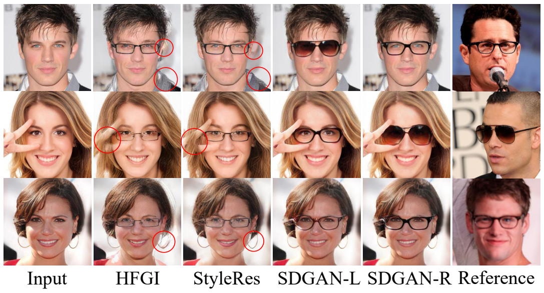


Figure 6: Qualitative comparison with HFGI and StyleRes on attribute "+ Eyeglasses". SDGAN-L: latent-guided; SDGAN-R: reference-guided.

Method	+ Eleglasses		- Eleglasses	
	↓ FID	↑ Acc	↓ FID	↑ Acc
HFGI	95.48	79.86	104.97	36.66
StyleRes	80.64	79.17	97.87	93.33
SDGAN-R	51.45	87.09	**75.17**	95.02
SDGAN-L	**49.39**	**98.52**	78.99	**95.33**

Table 4: Quantitative comparison with HFGI and StyleRes.

The quantitative results are shown in Table 4. It is worth mentioning that HFGI and StyleRes perform poorly in terms of FID because they are unable to manipulate target attribute styles to produce diverse outputs. On the other hand, our method excels in manipulating the target attribute style and consistently achieves the best FID and Acc scores.

Ablation Study

In this section, we conduct ablation studies to showcase the effectiveness of the proposed SDGAN. The qualitative results are presented in Fig.7. From the observations in Fig.7, we identify three crucial findings:

1) We replay the attribute-specific editing modules (i.e., T_i for different attributes in G) with an unified module (i.e., a single T) for all attributes editing as existing methods (w/o ASEM). This change affects other irrelevant content, e.g., hair color and illumination.
2) We remove the semantic masks and instead directly combine the outputs of different attribute-specific editing modules element-wise (w/o Mask). This change introduces noise (1st row) and color distribution (3rd row), and fails to preserve details of hair and beard (last row).

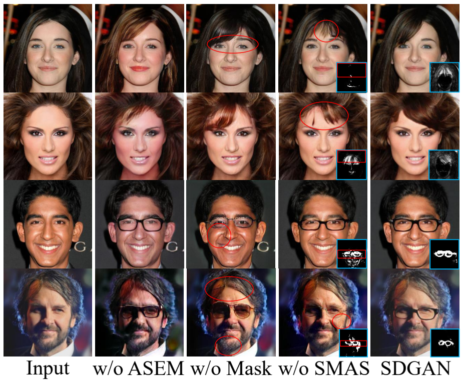


Figure 7: Qualitative results of the ablation study. The first two rows: + Bangs, the last two rows: + Eyeglasses.

3) We exclude the semantic mask alignment strategy (w/o SMAS). We provide the semantic mask corresponding to the target attribute in the bottom right corner of each image. Without SMAS, the semantic masks either focus on incorrect regions (1st row) or cover a much larger area than the actual region (3rd & 4th rows). Consequently, unrealistic editing results emerged when editing bangs, and undesirable changes occurred when editing eyeglasses. In contrast, SDGAN can focus on the relevant areas and suppress undesired changes.

Overall, SDGAN demonstrates high-quality editing results with correct modification and unrelated preservation. Moreover, SDGAN also consistently achieves the best quantitative performance, as listed in Table 3.

Conclusion

In this paper, we introduce a novel framework SDGAN to address the challenge of accurate and controllable facial attribute editing. By utilizing a semantic disentanglement generator with attribute-specific editing modules and semantic masks, we effectively separate the editing process for different attributes. Additionally, our novel semantic mask alignment strategy guides the semantic masks to precisely identify and restrict the editing regions. Extensive experiments demonstrate the superiority of SDGAN over state-of-the-art methods in style manipulation, image quality, and attribute editing accuracy. Moreover, we believe that the modules and strategies proposed in this paper can serve as valuable references for other related editing tasks, potentially enhancing their performance as well.

Acknowledgments

This work was supported by the National Science Foundation of China under Grants U2001202, U19B2022, and 61972430.

References

Alaluf, Y.; Tov, O.; Mokady, R.; Gal, R.; and Bermano, A. 2022. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *CVPR*, 18511–18521.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *ICLR*.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 8188–8197.

Dalva, Y.; Altındiş, S. F.; and Dundar, A. 2022. VecGAN: Image-to-Image Translation with Interpretable Latent Directions. In *ECCV*, 153–169.

Gafni, O.; and Wolf, L. 2020. Wish you were here: Context-aware human generation. In *CVPR*, 7840–7849.

Gao, Y.; Wei, F.; Bao, J.; Gu, S.; Chen, D.; Wen, F.; and Lian, Z. 2021. High-fidelity and arbitrary face editing. In *CVPR*, 16115–16124.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NIPS*, 2672–2680.

He, Z.; Zuo, W.; Kan, M.; Shan, S.; and Chen, X. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE TIP*, 28(11): 5464–5478.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 30.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *CVPR*, 1501–1510.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*, 1125–1134.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *CVPR*, 8110–8119.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 5549–5558.

Li, H.; Wang, W.; Yu, C.; and Zhang, S. 2021a. SwapInpaint: Identity-specific face inpainting with identity swapping. *IEEE TCSVT*, 32(7): 4271–4281.

Li, X.; Zhang, S.; Hu, J.; Cao, L.; Hong, X.; Mao, X.; Huang, F.; Wu, Y.; and Ji, R. 2021b. Image-to-image translation via hierarchical style disentanglement. In *CVPR*, 8639–8648.

Pehlivan, H.; Dalva, Y.; and Dundar, A. 2023. Styleres: Transforming the residuals for real image editing with stylegan. In *CVPR*, 1828–1837.

Petzka, H.; Fischer, A.; and Lukovnikov, D. 2018. On the regularization of Wasserstein GANs. In *ICLR*.

Shen, W.; and Liu, R. 2017. Learning residual images for face attribute manipulation. In *CVPR*, 4030–4038.

Shi, Y.; Yang, X.; Wan, Y.; and Shen, X. 2022. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In *CVPR*, 11254–11264.

Wang, T.; Zhang, Y.; Fan, Y.; Wang, J.; and Chen, Q. 2022. High-fidelity gan inversion for image attribute editing. In *CVPR*, 11379–11388.

Wang, Y.; Gonzalez-Garcia, A.; van De Weijer, J.; and Herranz, L. 2019. Sdit: Scalable and diverse cross-domain image translation. In *ACM MM*, 1267–1276.

Xia, M.; Shu, Y.; Wang, Y.; Lai, Y.-K.; Li, Q.; Wan, P.; Wang, Z.; and Liu, Y.-J. 2023. FEditNet: Few-shot Editing of Latent Semantics in GAN Spaces. In *AAAI*.

Xiao, T.; Hong, J.; and Ma, J. 2018. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *ECCV*, 168–184.

Xin, J.; Wang, N.; Jiang, X.; Li, J.; Gao, X.; and Li, Z. 2020. Facial attribute capsules for noise face super resolution. In *AAAI*, volume 34, 12476–12483.

Xu, Z.; Yu, X.; Hong, Z.; Zhu, Z.; Han, J.; Liu, J.; Ding, E.; and Bai, X. 2021. Facecontroller: Controllable attribute editing for face in the wild. In *AAAI*, volume 35, 3083–3091.

Yazici, Y.; Foo, C.-S.; Winkler, S.; Yap, K.-H.; Piliouras, G.; Chandrasekhar, V.; et al. 2019. The Unusual Effectiveness of Averaging in GAN Training. In *ICLR*.

Yin, W.; Liu, Z.; and Loy, C. C. 2019. Instance-level facial attributes transfer with geometry-aware flow. In *AAAI*, volume 33, 9111–9118.

Zhang, W.; Liu, Y.; Dong, C.; and Qiao, Y. 2019. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *ICCV*, 3096–3105.

Zhao, B.; Chang, B.; Jie, Z.; and Sigal, L. 2018. Modular generative adversarial networks. In *ECCV*, 150–165.